TITLE OF THE INVENTION:

CORRELATING VIDEO IMAGES OF LIP MOVEMENTS WITH AUDIO SIGNALS TO IMPROVE SPEECH RECOGNITION

CROSS REFERENCE TO RELATED APPLICATIONS:

[0001] This application claims priority of U.S. Provisional Patent Application Ser. Nos. 60/409,956, filed September 12, 2002, and 60/445,816, filed February 10, 2003, entitled Correlating Video Images of Lip Movements with Audio Signals to Improve Speech Recognition. The contents of the provisional applications are hereby incorporated by reference.

BACKGROUND OF THE INVENTION:

Field of the Invention:

[0002] The present invention relates to a method of and an apparatus for using video signals along with audio signals of speech to provide speech recognition, within an environment where speech recognition is necessary. In particular, the present invention relates to a method of and a system for using video images of lip movements with audio input signals to improve speech recognition. The present invention can be implemented in a hand held device, and the invention may include discrete devices or may be implemented on a semiconductor substrate such as a silicon chip.

Description of the Related Art:

[0003] Human speech is made up of numerous different sounds and syllables. Often in many languages, different sounds and/or syllables are combined to form words and/or sentences. The combination of the sounds, syllables, words, and sentences forms the basis for oral communication.

[0004] Generally, human speech is recognizable if the speech is clear and comprehensible to another human's ears. On the other hand, human speech can be recognizable by a machine if the audio waves of the speech is received, and the audio waves are recognizable by an algorithm operating within the machine. Although audio speech recognition by machines has

advanced in sophistication, the accuracy of audio speech recognition has room for improvements.

SUMMARY OF THE INVENTION:

[0005] One example of the present invention can be a method of speech recognition. The method can include the steps of receiving audio signals from a speech source, receiving video signals from the speech source, and processing the audio signals and the video signals. The method can also include the steps of converting the audio signals and the video signals to recognizable information, and implementing a task based on the recognizable information.

[0006] In another example, the present invention can relate to a speech recognition device. The device can have an audio signal receiver configured to receive audio signals from a speech source, a video signal receiver configured to receive video signals from the speech source, and a processing unit configured to process the audio signals and the video signals. Moreover, the device can have a conversion unit configured to convert the audio signals and the video signals to recognizable information, and an implementation unit configured to implement a task based on the recognizable information.

[0007] Additionally, another example of the present invention can provide a system for speech recognition. The system can include a first receiving means for receiving audio signals from a speech source, a second receiving means for receiving video signals from the speech source, and a processing means for processing the audio signals and the video signals. Furthermore, the system can have a converting means for converting the audio signals and the video signals to recognizable information, and an implementing means for implementing a task based on the recognizable information.

[0008] Furthermore, another example of the present invention can be directed to a method of speech recognition. The method can include the steps of receiving audio signals from a speech source, receiving video

signals from the speech source, processing the audio signals, and converting the audio signals into recognizable information. Moreover, the method can have the step of processing the video signals when a segment of the audio signals can not be converted into the recognizable information. The video signals can coincide with the segment of the audio signals that cannot be converted into the recognizable information. The method also can have the steps of converting the processed video signals into the recognizable information, and implementing a task based on the recognizable information.

[0009] In another example, the present invention can be a speech recognition device. The device can have an audio signal receiver configured to receive audio signals from a speech source, a video signal receiver configured to receive video signals from the speech source, a first processing unit configured to process the audio signals, and a first conversion unit configured to convert the audio signals to recognizable information. The device can also have a second processing unit configured to process the video signals when the audio signals cannot be converted into the recognizable information, wherein the video signals coincide with the segment of the audio signals that cannot be converted into the recognizable information, a second conversion unit configured to convert the video signals processed into the recognizable information, and an implementation unit configured to implement a task based on the recognizable information.

[0010] In yet another example, the present invention can be drawn to a system for speech recognition. The system can include a first receiving means for receiving audio signals from a speech source, a second receiving means for receiving video signals from the speech source, a first processing means for processing the audio signals, and a first converting means for converting the audio signals into recognizable information. The system can also have a second processing means for processing the video signals when a segment of the audio signals can not be converted into the recognizable information, wherein the video signals coincide with the segment of the

audio signals that cannot be converted into the recognizable information, a second converting means for converting the video signals processed into the recognizable information, and an implementing means for implementing a task based on the recognizable information.

BRIEF DESCRIPTION OF THE DRAWINGS:

[0011] For proper understanding of the invention, reference should be made to the accompanying drawings, wherein:

[0012] Figure 1 illustrates one example of a speech recognition device using audio signals and correlating video signals to improve speech recognition, in accordance with the present invention;

[0013] Figure 2 illustrates a flow chart illustrating one example of a method of speech recognition using audio signals and correlating video signals, in accordance with the present invention;

[0014] Figure 3 illustrates a flow chart illustrating another example of a method of speech recognition using audio signals and correlating video signals, in accordance with the present invention;

[0015] Figure 4 illustrates a flow chart of a method of speech recognition using audio signals and correlating video signals, in accordance with the present invention; and

[0016] Figure 5 illustrates one example of a hardware configuration for speech recognition using audio signals and correlating video signals, in accordance with the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT(S):

[0017] Figure 1 illustrates one example of a speech recognition device for improving speech recognition with audio input signals and correlating video input signals according to the present invention. Figure 1 shows a mobile phone 100 having a display screen 101 and a plurality of actuators 102. In addition, the mobile phone 100 can include a lens 103 for receiving images in video and/or still picture format. In the example shown in Figure 1, the

lens 103 can capture or receive video images of the movements of a user's lips 104a, 104b made while the user is speaking.

[0018] For instance, a user may desire to contact a business associate via an e-mail message. According, the user can access the mobile phone 100 and can activate the speech recognition system of the present invention. The mobile phone 100 is placed adjacent to the user's face where the lens 103 is positioned in proximity to the user's mouth so that the image of the user's lips can be captured by the speech recognition system. Once the lens 103 is positioned correctly, the user can be alerted to commence speaking into the mobile phone 100. The user, for example, can speak and request to send an e-mail message to Jane Doe in which her e-mail address is pre-programmed into the mobile phone 100. In addition, the name James Doe is similarly pre-programmed in the mobile phone 100. The speech recognition system processes the audio signal input and can convert all part of the audio signal input into recognizable information with the exception of the name Jane Doe. The audio speech recognition feature of the invention cannot ascertain if the audio signal is referring to Jane Doe or James Doe.

[0019] Accordingly, the speech recognition system of the present invention can access a section of the video input signal corresponding to the audio signal pertaining to Jane Doe. Based on the detected lip movements of the video signals, the present invention can reduce the uncertainty of the audio speech recognition, and therefore can perform speech recognition with the aid of the video signals, and determine that the request is an e-mail message to Jane Doe rather than James Doe. Thereafter, the present invention can implement one or more actions to carry out the spoken request by the user. The one or more actions can be in the forms of commands such as initiating an e-mail application software, creating an outgoing e-mail window, and inserting text, and sending the e-mail message.

[0020] Although the example provided in Figure 1 illustrates a mobile phone 100 having a lens 103, wherein the mobile phone 100 can be

configured with the speech recognition system of the present invention, it is noted that the speech recognition system using audio signals and correlating video signals of the invention can be configured on a variety of electronic device, either mobile or stationary. For instance, the improved speech recognition system of the invention can be configured on at least but not limited to a laptop computer, a PDA, an audio/video recording device, a home computer, a game console, a remote controller, or other comparable device.

[0021] Figure 2 illustrates one example of a method of speech recognition using audio input signal and correlating video input signals, in accordance with the present invention. Specifically, Figure 2 illustrates one example of a method of speech recognition using audio input signals together with correlating video images of lip movements. The method of the present example can be implemented in hardware, or software, or a combination of both hardware and software.

[0022] Figure 2 illustrates one example of a method of speech recognition according to the present invention. A device configured to include a speech recognition system can be activated at step 200 of Figure 2. In other words, the present invention provides a user with the option to activate the speech recognition feature when necessary. After the speech recognition system is activated, a detecting sensor along with an optical pick-up such as a lens, can detect for video images resembling the user's lips at step 201. If the detecting sensor and the lens do not detect images of the user's lips, then the speech recognition system can alert the user to readjust the lens or the device, or reposition the user's lips so that an image can be detected, at step 202 of Figure 2. If however, the user's lips can be detected or captured by the lens and the sensor, then the speech recognition system can alert the speaker in step 203 that the speech recognition system is in ready mode, and therefore the user can commence speaking.

[0023] Once the user starts to speak, the speech recognition system of the present invention can commence receiving both audio and video input signals from the user's speech and the user's lip movements at step 204. In this example, the speech recognition system can process the audio input signals corresponding to the speech first. In other words, as the user speaks, both the audio speech and the correlating images of the user's lip movements can be received by the speech recognition system of a device. Although both audio and video signals are being received, the speech recognition system can preliminarily initiate only the audio speech recognition portion of the system, and can preliminarily process only the audio portion of the speech.

[0024] Therefore, if the speech from the user does not contain a possibly unrecognizable sound or word, then the present invention at step 206 can recognize the speech as comprehensible and recognizable information using only the audio speech recognition portion of the system without the need to activate the assistance of the video signals.

[0025] The speech recognition system can process the audio input signals and determine if the audio input signals are recognizable as speech at step 205. If it is determined that the audio input signals corresponding to a user's entire speech can be processed and converted into recognizable information, then the speech recognition system can process the entire audio input signals and convert it to recognizable information at step 206, without initiating the video signal speech recognition functions.

[0026] Thereafter, the speech recognition system can implement one or more task(s) based on the recognizable information at step 207. For instance, the speaker can talk into a cell phone configured with the speech recognition system of the present invention. The speaker can request to dial a particular number or connect with the Internet. Therefore, the speech recognition system can convert the speech into either recognizable information such as numeric characters like dialing a particular number, or convert the speech into recognizable information such as a set of code(s) to

perform a particular function like connecting with the Internet. Accordingly, the audio signal speech recognition processing functions can become the primary processing functions of the speech recognition system until a section of the audio input signals cannot be processed and converted into recognizable information.

[0027] If however a section of the audio input signals cannot be process and converted into recognizable information, the present invention can access the correlating portion of the video input signals at step 208 to assist in recognizing the speech. In other words, whenever the audio speech recognition portion of the system identifies a possibly unrecognizable sound or word, then the speech recognition system can access a portion of the video image of the lip movements of the speaker, wherein the video image can correspond to the unrecognizable audio signal portion. For instance, audio input signals based on a user's speech can be received by the speech recognition system, and when the audio speech recognition portion of the system detects a possible conversion error that is equal to or is above a predetermined threshold level, then the video speech recognition portion of the system can be initiated.

[0028] Once the video speech recognition portion of the system is initiated, the system can access the video images of the lip movements correlating to the audio speech in question and can determine the movements of the lips at step 209. The system can thereafter process the video images and assist in the conversion of audio and video input signals to recognizable and comprehensible information at step 210. It is noted that although the video input signals can be processed to assist in speech recognition, the video input signal can also be processed not just as an aid to the audio input signals but as a stand-alone speech recognition feature of the system.

[0029] Following the processing and converting of the video input signals correlating to the audio signal in question, the speech recognition system can implement a task based on the recognizable information at step 211.

[0030] Thus, the combination of both the audio speech and the video image of the lip movement can resolve unrecognizable audio speech. In addition, the system can be configured to identify likely sounds corresponding to certain lip movements, and can also be configured to recognize speech based on the context in which the word or sound was spoken. In other words, the present invention can resolve unrecognizable speech by referring to the adjacent recognizable words or phrases within the speech in order to aid in the recognition of the unrecognizable portion of the speech.

[0031] In an alternative example, the present invention can recognize speech using both audio and video signals at a destination site rather than at the originator. Figures 3 and 4 illustrates one example of a method of sending the audio and video input signals to a destination site where the audio and video input signals can be processed and converted into recognizable and comprehensible information. Specifically, Figures 3 and 4 illustrate one example of a method of speech recognition at a destination site using audio input signals together with correlating video images of lip movements. The method of the present example can be implemented in hardware, or software, or a combination of both hardware and software.

[0032] A device configured to include a speech recognition system can be activated at step 300 of Figure 3. After the speech recognition system is activated, a detecting sensor along with an optical pick-up such as a lens, can detect for video images resembling the user's lips at step 301. If the detecting sensor and the lens do not detect images of the user's lips, then the speech recognition system can alert the user to readjust the lens or the device, or reposition the user's lips so that an image can be detected, at step 302 of Figure 3. If however, the user's lips can be detected or captured by the lens and the sensor, then the speech recognition system can alert the speaker in step 303 that the speech recognition system is in ready mode, and therefore the user can commence speaking.

[0033] Once the user starts to speak, the speech recognition system of the present invention can commence receiving both audio and video input signals from the user's speech and the user's lip movements at step 304. The received audio input signals and the received video input signals can be stored within a storage unit and/or a plurality of separate storage units.

[0034] Following the completion of the user's speech, the speech recognition system can detect, based on sensors and preprogrammed conditions, an end of speech status at step 305. In other word, once the sensors detects that the user has completed his speech and that certain preprogrammed conditions have been met, then speech recognition system can activate an end of speech condition. Thereafter, the speech recognition system can prompt the user if the user desires to send the stored speech at step 306.

[0035] If the user responds in the negative, then the stored speech can remain stored in the storage unit(s) and can be recalled at later time. However, if the user responds in the positive, then the speech recognition system can transmit the stored speech to a destination site at step 307.

[0036] After the stored speech is received at the destination site, then the destination site can activate the audio and video speech recognition system available at the destination site at step 400. Thereafter, the audio and video speech recognition system can process and convert the audio and video signals to recognizable and comprehensible information as discussed above with respect to Figure 2. In other words, the speech recognition system at the destination site can preliminary determine if the audio input signal can be processed and converted to recognizable information at step 401. If the entire audio portion of the speech or the entire audio input signals can be processed and converted as recognizable information, then the system can do so without activating the video speech recognition portion of the system at step 406. Thus, the entire audio portion of the speech can be processed and converted into recognizable information, and the speech recognition system

can implement one or more task(s) based on the recognizable information at step 407.

[0037] If, however, a section of the audio input signals cannot be process and converted into recognizable information, the present invention can access the correlating portion of the video input signals at step 402 to assist in recognizing the speech. In other words, whenever the audio speech recognition portion of the system detects a possibly unrecognizable sound or word, then the speech recognition system can access a portion of the video image of the lip movements of the speaker, wherein the video image can correspond to the unrecognizable audio signal portion.

[0038] Once the video speech recognition portion of the system is triggered, the system can access the video images of the lip movements correlating to the audio speech in question and can process and determine the movements of the lips at step 403. The system can thereafter process the video images and assist in the conversion of audio and video input signals to recognizable and comprehensible information at step 404. After the conversion of the audio and/or video input signals, the speech recognition system can implement one or more task(s) based on the recognizable information at step 405.

[0039] It is noted that the speech recognition system of the present invention can simultaneously process and convert the audio input signals and the video input signals in parallel. In other words, rather than the system initiating the audio speech recognition portion of the system to first process and convert the audio portion of the speech, the system can initiate both the audio speech recognition in tandem with the video speech recognition. Therefore, the speech recognition system can process the audio input signals and the correlating video input signals in parallel, and can convert the audio speech and the correlating video images of the lip movements into recognizable and comprehensible information.

[0040] Figure 5 illustrates one example of a hardware configuration that can perform speech recognition based on audio input signals and correlating video input signals, in accordance with the present invention. In addition, the hardware configuration of Figure 5 can be in an integrated, modular and single chip solution, and therefore can be embodied on a semiconductor substrate, such as silicon. Alternatively, the hardware configuration of Figure 5 can be a plurality of discrete components on a circuit board. The configuration can also be implemented as a general purpose device configured to implement the invention with software.

[0041] Figure 5 illustrates a device 500 configured to perform speech recognition based on audio signals and correlating video images of lip movements. Device 500 can contain an audio receiving unit 505 and a video receiving unit 510. The audio receiving unit 505 can receive audio input signals from one or more audio source(s) such as voice, speech, music, etc. The video receiving unit 510 can receive video input signals from one or more video source(s). For example, the video receiving unit 510 can receive video images of a speaker's lip movements. In addition, the device 500 can include a video image sensor 515, wherein the sensor 515 can detect when a particular image such as a speaker's lips is not being received by the video receiving unit 510. In other words, if the speaker's lips are not positioned in a way for the video receiving unit to receive video images of lips, then the sensor can detect missing video images and can alert the speaker.

[0042] Furthermore, the device 500 can include a processing unit 520 and a converting unit 525. The processing unit 520 can process the audio input signals as well as the video input signals. The converting unit 525 can convert the processed audio input signals and the video input signals into recognizable and comprehensible information. For instance, the converting unit 525 can convert the processed audio input signals and the video images of lip movements into executable commands or into text, etc. If the

converted signals are commands to perform one or a set of function(s), then the implementation unit 530 can execute the command(s).

[0043] One having ordinary skill in the art will readily understand that the invention as discussed above may be practiced with steps in a different order, and/or with hardware elements in configurations which are different than those which are disclosed. Therefore, although the invention has been described based upon these preferred embodiments, it would be apparent to those of skill in the art that certain modifications, variations, and alternative constructions would be apparent, while remaining within the spirit and scope of the invention. In order to determine the metes and bounds of the invention, therefore, reference should be made to the appended claims.